DOCUMENT RESUME

ED 262 091 TM 850 572

AUTHOR Jansen, Hans P.
TITLE Training Emphasis Task Factor Data: Methods of Analysis.
INSTITUTION Air Force Human Resources Lab., Brooks AFB, Tex. Manpower and Personnel Div.
REPORT NO AFHRL-TR-84-50
PUB DATE May 85
NOTE 26p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cluster Analysis; Computer Software; Evaluation Methods; *Factor Analysis; *Instructional Development; *Interrater Reliability; *Job Analysis; Measurement Techniques; *Military Training; Needs Assessment; Rating Scales; Sample Size; Skill Analysis
IDENTIFIERS Air Force; *REXALL (Computer Software)

ABSTRACT

The Air Force Occupational Measurement Center conducts task-based occupational surveys of Air Force specialties that include supervisor ratings on recommended training emphasis for entry-level airmen. Priorities are input to the Instructional System Development training model, which guides the development and revision of specialty training courses. For 20 percent of specialties, training emphasis ratings have been subject to poor interrater agreement. Data may contain conflicting rating policies within a specialty. To develop a methodology for identifying multiple rating policies in such data, this research investigates: (1) the variation in interrater agreement with respect to sample size; and (2) the multiple rating policy hypothesis via modified REXALL analysis, cluster analysis, and factor analysis. Agreement is found to vary within and across sample sizes, and a minimum of 55 raters is recommended. REXALL analyses are inconclusive with respect to confirming the presence or absence of multiple rating policies. Results indicate that samples of training emphasis ratings are less complex than expected. REXALL analyses are recommended for single ladder specialties; principal components factors analysis with VARIMAX rotation is recommended for multiple factors--extracting one and then multiple factors as appropriate. Interpretation of these results can be enhanced with CODAP auxiliary programs. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AFHRL-TR-84-50

AIR FORCE HUMAN RESOURCES LABORATORY

ED262091

TRAINING EMPHASIS TASK FACTOR DATA: METHODS OF ANALYSIS

By

Hans P. Jansen
Squadron Leader, Royal Australian Air Force Exchange Officer

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

May 1985
Final Report

Approved for public release; distribution unlimited.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
AFHRL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 572

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

BEST COPY AVAILABLE

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

ANTHONY F. BRONZO, JR., Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AFHRL-TR-84-50	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Manpower and Personnel Division 6b. OFFICE SYMBOL (*If applicable*) AFHRL/MO	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (*City, State and ZIP Code*) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601	7b. ADDRESS (*City, State and ZIP Code*)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (*If applicable*) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (*City, State and ZIP Code*) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
62703F	7719	19	11, 01
62703F	7734	07	30

11. TITLE (*Include Security Classification*) Training Emphasis Task Factor Data: Methods of Analysis			
12. PERSONAL AUTHOR(S) Jansen, Hans P.			
13a. TYPE OF REPORT Final	13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (*Yr., Mo., Day*) May 1985	15. PAGE COUNT 28
16. SUPPLEMENTARY NOTATION			

17. COSATI CODES			18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB. GR.	common rating policy; Comprehensive Occupational Data Analysis Programs (CODAP); interrater reliability; rating policies; REXALL computer program

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

REXALL, a program within the Comprehensive Occupational Data Analysis Programs (CODAP) system, is routinely used for assessing the level of interrater agreement obtained when multiple raters evaluate "training emphasis" at the task level and for extracting a reliable common rating policy (CRP). For some samples, very poor interrater agreement precludes extraction of a reliable CRP and limits use of the data. Since poor interrater agreement may be a function of multiple rating policies, research was initiated to develop a methodology for identifying the multiple rating perceptions that may exist within task factor data. The findings presented include the effect of sample size on interrater agreement and the use of modified REXALL analysis, cluster analysis and factor analysis techniques for identifying multiple rating policies in training emphasis data. Results indicate that REXALL analysis employing new CRP extraction criteria is adequate for samples where the CRP includes all raters and when the CRP has a divergency of less than 25%. It was also found that principal components factor analysis has high utility for identifying the CRP and any other rating policies that might exist in the rating data. Possible causes of poor interrater agreement and several alternative approaches to identifying the causes for interrater disagreement are discussed. Guidelines for occupational analysts to follow when using REXALL and alternative analysis procedures are provided.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION	
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy A. Perrigo Chief, STINFO Office	22b. TELEPHONE NUMBER (*Include Area Code*) (512) 536-3877	22c. OFFICE SYMBOL AFHRL/TSR

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

SUMMARY

The Air Force Occupational Measurement Center (USAFOMC) conducts occupational surveys of Air Force specialties. These include the collection of supervisors' ratings on task factors such as recommended emphasis for first-term training. The task training emphasis ratings serve as input to the Instructional System Development (ISD) training model, which guides the development and revision of technical training courses. Analysis of training emphasis ratings is usually performed using REXALL, a special-purpose program within the Comprehensive Occupational Data Analysis Programs (CODAP) system. Two important functions of REXALL are to assess the overall level of agreement among raters, and to calculate an average (mean) factor rating for each task. When an acceptable level of interrater agreement is attained, the task means are rank-ordered. This rank-ordering constitutes the recommended priority of training for each of the tasks and defines the common rating policy (CRP) for the specialty.

For a small number of specialties, referred to as "complex specialties," very poor interrater agreement is frequently found that precludes the extraction of a reliable training emphasis CRP. Driven by the suggestion that poor interrater agreement may be caused by competing rating policies with possible relevance to training, a Request for Personnel Research (RPR) was initiated by USAF OMC and validated through Hq Air Training Command. The RPR requested development of a methodology for identifying multiple rating policies that might exist in such data.

Research on the possible causes of poor interrater agreement followed two main courses: (a) investigation of the variation in interrater agreement with respect to the number of raters used (sample size) and (b) investigation of the multiple-rating-policy hypothesis via three independent analysis techniques: modified REXALL analysis, cluster analysis, and factor analysis. These techniques were applied to seven "complex specialties" to see if multiple rating policies could be identified.

Interrater agreement was found to vary within and across different sample sizes. A sample of approximately 55 raters is the minimum number recommended for extraction of a reliable CRP. REXALL analyses were inconclusive with respect to confirming the presence or absence of multiple rating policies. Cluster analyses using existing CODAP software also proved to be generally inadequate for identifying multiple rating policies. However, some CODAP programs that report rater responses in clustering (KPATH) sequence were found to be highly useful for interpreting observed REXALL statistics.

Results of principal components factor analyses clearly demonstrated that the samples of training emphasis ratings were less complex than expected. A one-factor solution confirmed that REXALL analyses which employ modified CRP extraction criteria are appropriate and sufficient for single-specialty samples which contain a dominant CRP. Where such REXALL analysis failed, additional analysis using a VARIMAX rotation/factor-building methodology successfully isolated significantly different multiple rating policies.

It is recommended that REXALL analyses with modified CRP extraction criteria be used for the vast majority of single-ladder specialties, where one might expect a single dominant training policy. In those cases when evidence suggests that multiple policies might be operative, principal components factors analysis with VARIMAX rotation is recommended--extracting one and then multiple factors as appropriate. Intepretation of these results can be enhanced with CODAP auxiliary programs (DUVARS, PRTDIS, PRTVAR and FACPRT).

BEST COPY AVAILABLE

Preface

This work resulted from Request for Personnel Research (RPR) 79-1, Analysis of Ratings by Occupational Task Factors; from Headquarters Air Training Command, and was initiated under Work Unit 77340730, Complex Specialties Task Training Priority Equation Development. It was subsequently completed under Work Unit 77191911, Measurement and Analysis of Job and Mission Requirements. The present effort represents a portion of the Laboratory's Force Acquisition and Distribution System thrust.

Dr. William Alley and Dr. Hendrick Ruck provided helpful suggestions and significant assistance in the conduct of this effort.

TABLE OF CONTENTS

		Page
I.	BACKGROUND	5
II.	FINDINGS	8
	Sampling Variations	8
	Detecting Multiple Rating Policies	11
	Modified REXALL Analysis	11
	Cluster Analysis	13
	Factor Analysis	14
III.	APPLICATIONS	19
IV.	CONCLUSIONS	21
REFERENCES		21
Appendix A:	CODAP Clustering Description	23

LIST OF FIGURES

Figure		
1	Single-specialty rating policy domain	8
2	Stability of R_{11} versus sample size	11

LIST OF TABLES

Table

1 Training Emphasis Data Samples Analyzed With All Raters Included 9

2 Training Emphasis Data Analyzed With Sequential Removal of Divergent Raters. 10

3 Variation in R_{jj} with Sample Size. 10

4 Frequency of Occurrence of Rater Correlations. 12

5 Percentage of Occurrence of Rater Correlations 13

6 Analysis Results for the General Factor (CRP) for Each Specialty 15

7 Comparison of General Factor (CRP) and Second Iteration Deletion Statistics for Each Specialty. 16

8 General and Rated Factor Statistics for AFSC 404X0 18

9 Rotated Factor Solution for AFSC 404X0 18

10 Rotated Factor Solution for AFSC 328XX 19

TRAINING EMPHASIS TASK FACTOR DATA: METHODS OF ANALYSIS

I. BACKGROUND

The Air Force Occupational Measurement Center (USAFOMC) conducts task-based occupational surveys of Air Force specialties. These surveys include the collection of supervisors' ratings on task factors such as recommended training emphasis. Recommended training emphasis is defined as the emphasis that should be given in structured training of the task for entry-level airmen, regardless of where that training takes place (i.e., resident course, Field Training Detachment, or on-the-job training). First-term training priorities are input to the Instructional System Development (ISD) training model, which guides the development and revision of specialty training courses. The utility, reliability, and validity of training emphasis ratings in terms of ISD theory have been demonstrated by Ruck, Thompson, Brown, and Stacy (in preparation).

For approximately 20% of specialties, training emphasis ratings have been quite difficult to interpret, due to poor interrater agreement. The suggestion has been that the data for such a "complex specialty" may contain conflicting rating policies aligned with the various employment duties/areas within a specialty. Currently, there are no satisfactory operational techniques for identifying such multiple policies. Research to develop a methodology for identifying the various rating perceptions that may exist in training emphasis ratings was initiated as a result of a Request for Personnel Research (RPR 79-1), Analysis of Ratings by Occupational Task Factors, submitted by Headquarters Air Training Command.

Analysis of training emphasis rating data is usually performed using REXALL, a special-purpose program developed and documented by Christal and Weissmuller (1976) within the Comprehensive Occupational Data Analysis Programs (CODAP) system. The three main functions of REXALL are (a) to assess the level of interrater agreement, (b) to identify divergent raters, and (c) to calculate the mean factor rating for each task. With respect to overall interrater agreement, REXALL is designed to cope with a sample of raters who are anticipated to be relatively homogeneous in terms of their rating ability.

Ratings for first-term training emphasis are made using a 9-point scale: from 1 (extremely low) to 9 (extremely high). However, the instruction to "rate only tasks which you believe require training for first-termers" recognizes the validity of a zero rating. By default, all non-ratings are interpreted to mean "no training recommended" and are included as zeros in all REXALL calculations, including the mean training emphasis for each task.

As a measure of interrater agreement, REXALL computes two indices of interrater reliability using the intraclass correlation formulas reported by Lindquist (1953). The two indices are R_{11}, single-rater reliability, which approximates the average of all possible pair-wise rater correlations; and R_{kk}, reliability for a sample of k raters, which is the expected correlation between the set of observed sample task means and the task means of an hypothetical equivalent sample. R_{11}'s and R_{kk}'s meeting or exceeding minimum criterion values are interpreted as meaning that sufficient interrater agreement exists to produce stable estimates of task mean values.

The standard REXALL analysis procedure for achieving acceptable interrater agreement and a set of reliable task mean ratings is to identify and delete divergent raters, as discussed by Goody (1976). Divergent raters are those whose ratings differ significantly from the ratings of the majority of raters because of failure to follow instructions, inverted or poor discriminative use of the rating scale, unique perception of tasks, or lack of knowledge. These divergent rater characteristics are reflected by a low or negative correlation between the individual rater's set of ratings and the sample task means (excluding the subject rater's ratings), and/or a low

t-value (confidence level associated with the correlation being different from zero). A typical rater sample is assumed to have a simple structure consisting of a majority of good raters who yield a set of stable task means and a minority of divergent raters who individually disagree with the majority rating pattern. For determining training emphasis, the rank-ordered task means computed from the ratings of the residual good raters constitute the recommended training priority and define the common rating policy (CRP).

The REXALL program provides no information as to why, for some specialties, R_{11} remains low even after successive deletions of divergent raters. The rationale underlying the present effort is that for such specialties, a low R_{11} may be a function of conflicting multiple rating policies, each associated with a subgroup of raters sharing similar training perceptions aligned with a specific employment area within the specialty. If this is the case, then the mean ratings, across a total specialty sample, may not reflect any meaningful policy, and significant policy differences may be obscured by the averaging process.

The present study was aimed at developing a technique to identify and describe such different policies which, when present, may account for the low interrater reliabilities obtained for some specialties. In designing the approach, it was recognized that other factors may also contribute to low interrater agreement. Five factors, in all, were regarded as possible sources of error: (a) random sampling variance, (b) multi-ladder task lists, (c) random variation in rater responses, (d) presence of divergent raters, and (e) multiple rating policies. The first of these, random sampling variance, was investigated by observing the effects on R_{11} of repeated samplings involving different numbers of raters. The remaining factors were investigated employing modified REXALL analysis, CODAP cluster analysis, and factor analysis. These techniques are described under "Findings." The paragraphs that follow discuss five possible causes of low R_{11}.

1. Random sampling variance, a function of sample size, was considered to be a potentially significant cause of low interrater agreement. The average operational training emphasis sample size is 45 supervisory raters, with a range of 10 to 80 raters. The sample size is primarily a function of supervisory rater availability. Statistically, there is a greater chance of obtaining an unrepresentative sample with abnormally low (or high) interrater agreement for the smaller samples. The relationship between sample size and the interrater reliability indices, R_{11} and R_{kk}, is algebraically summarized by the Spearman-Brown prophecy formula. In general terms, it states that R_{kk} increases as R_{11} and sample size increase. The criterion minimum for acceptable single rater reliability, $R_{11} = .20$, is obtained from this formula by the insertion of $R_{kk} = .90$ as a widely recognized criterion minimum for stable task means, and a sample size of approximately 40 raters which is regarded as sufficiently large to be stable. Estimation of this minimum sample size assumes the level of interrater agreement and basis for agreement (rating policy) within the sample reflects that of the parent population. To address the issue of the stability of R_{11} as a function of sample size, two large, single-specialty rater samples were taken as independent finite populations, and 100 subsamples for each of 12 sample-size points in the 10- to 100-rater range were randomly selected and assessed for level of single-rater reliability (R_{11}). The results are provided in the "Findings" section of this report.

2. Where more than one specialty is surveyed with a single comprehensive survey instrument (i.e., for multi-ladder task lists), a low R_{11} may be attributable to conflicting specialty-aligned interests with little or no common training recommended. REXALL analysis would obviously be inappropriate under this condition. Analysis results of a dual-specialty sample, both in combined form and as two single specialties, are included in the investigation of multiple rating policies.

3. Random variation in rater responses may occur where most raters disagree due to their highly individual interpretations of the task list and/or rating scale. This represents the extreme multiple-rating-policy condition. Although the research approach taken here uses cluster and factor analyses as primary methods, an understanding of how interrater agreement is assessed, and how rating policies are examined using existing techniques is in order. Being the primary ratings analysis tool readily available in CODAP, REXALL is normally used for analyses of all ratings.

4. The presence of divergent raters may serve to depress interrater agreement. Existing REXALL procedures for extracting a reliable CRP involve the initial deletion of the divergent raters (pass 1) and, if necessary, deletion of any newly identified divergent raters (pass 2). Divergent raters are eliminated from the sample to achieve stable estimates of task means. Consistently observed increases in R_{jj} and R_{kk} resulting from the deletion of divergent raters in operational samples support this procedure and contribute to the face validity of the following USAFOMC CRP extraction criteria for training emphasis: (a) minimum acceptable level of interrater agreement, $R_{jj} = .20$, $R_{kk} = .90$; (b) minimum acceptable rater correlation with mean, $\underline{r} = .30$ and/or $\underline{t}$-value = 3.0; (c) deletion boundaries - maximum of two deletion passes, maximum of 10% raters deleted; and (d) minimum number of good raters, 40. Complex specialties are defined as those whose training emphasis ratings fail to provide a reliable CRP via application of these procedures and criteria. However, the presence of an inordinate number of divergent raters may disguise an underlying CRP to an extent which renders existing CRP extraction criteria unsuitable. If, on the other hand, excessive rater divergence is viewed not as a distinction between good and poor raters, but as an indicator of multiple rating policies, then the fifth factor comes into play. This factor assumes the adequacy of the listed CRP extraction criteria for small or moderate divergence and assumes complexity to be attributable to competing rating policies when interrater agreement and divergence criteria are not met. It is important to note that the multiple rating policy condition does not preclude the possibility of a CRP which is not readily discernible via standard REXALL analysis nor the existence of divergent raters.

5. Multiple rating policies can be defined in terms of differences in the rank-ordering of tasks between various paired subgroups of raters. A Spearman rank-order correlation with an $\underline{r}_s < .50$ was taken as indicating a practical difference in the recommended training priority between any two rating policy groups. These differences may be attributed to any combination of differences in number, type, and level of tasks recommended. The greatest possible difference between any two policies is that they recommend totally different sets of tasks for training. Relatively small policy differences would result from minor variation in the level of recommendations on the same set of tasks. In relation to meaningful alternative training policies, it would be highly desirable for raters within significantly different rating policy groups to share a common background characteristic such as job title or major command (MAJCOM), which could be viewed as explanatory factors contributing to policy differences.

The postulated single-specialty rating policy domain is summarized in Figure 1. The simple or complex specialty classification corresponds to achievement or nonachievement of a reliable CRP employing the previously described standard REXALL analysis procedure and criteria. The multi-ladder sample type is not included in Figure 1 since this type is obviously predisposed to being complex and is, therefore, unsuitable for REXALL analysis.

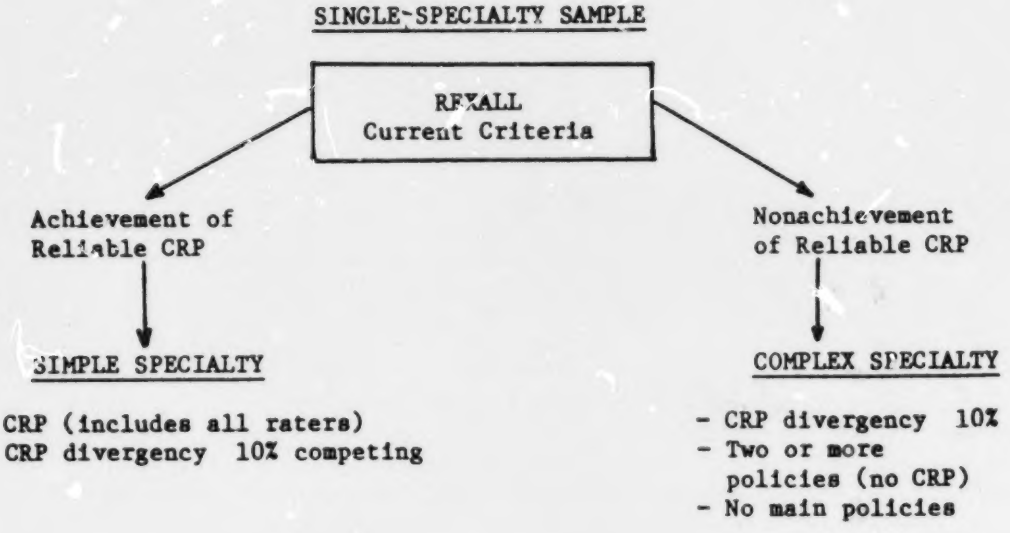

Figure 1. Single-specialty rating policy domain.

In the current investigation various analytical techniques were tested with training emphasis data from six specialties. Details for the six training emphasis data sets analyzed in this study are summarized in Table 1. The first two data sets were obtained from USAFOMC as examples of complex specialties with very poor interrater agreement. The third USAFOMC data set, a two-career-ladder study, was analyzed both in the combined form and as two single-specialty samples. The remaining two data sets were for specialties deemed complex as a consequence of the AFHRL training emphasis equation study (Ruck et al., in preparation). Application of standard criteria for deletion of divergent raters produces levels of interrater agreement as per Table 2. All samples fail to qualify as a simple specialty under strict application of the 10% maximum deletion criterion. However, the relatively high levels of interrater agreement for AFSCs 328X0, 328X1, and 672X2 suggest the specialties to be simple rather than complex. Attainment of minimum interrater agreement with a relatively high deletion percentage for AFSCs 811X0 and 304X0 render them possible complex specialties. The small AFSC 404X0 sample and the dual-specialty AFSC 328XX sample are complex.

II. FINDINGS

The findings presented pertain to the investigations of sampling error and multiple rating policies as possible causes of poor interrater agreement.

Sampling Variations

Two specialties, 304X4 and 672X2, were selected as probable complex specialties and rating data were collected from especially large samples of raters to permit analysis of sample size effects. Table 3 details the variation in R_{jj} at three sample sizes (10, 50, and 100 raters) for the two specialties. In each case, the average R_{jj} ($\bar{X}$) and variation in R_{jj} (SD) are for 100 random subsamples. The observed range in R_{jj} is described by the MIN and MAX values which illustrate the extent to which observed interrater agreement differed from that of the parent

population for a typical operational sample of 10 to 100 raters. The relationship between the stability of R_{11} (SD of R_{11}) and sample size is graphically summarized by the curves through the data points in Figure 2. Both Table 3 and Figure 2 demonstrate that, for corresponding sample sizes, the variation in R_{11} for the AFSC 672X2 raters is greater than that for the AFSC 304X4 raters with stabilization of R_{11} (SD = .02) occurring at n = 100 and n = 50, respectively. With respect to establishing a suitable sample size for REXALL analysis, both specialties are sufficiently stable at the 50- to 60-rater size to permit extraction of the CRP (if present). For sample sizes much below 50 raters, the problem of sampling error, as a cause of poor interrater agreement, is more significant.

Table 1. Training Emphasis Data Samples Analyzed with All Raters Included

AFSC	Title	Source	Number		R_{11}	R_{kk}
			Raters	Divergents		
404X0	Precision Imagery and Audio-Visual Media Maintenance	USAFOMC	47	12	.09	.73
811X0	Security Specialist	USAFOMC	120	23	.15	.95
328XX	Avionics Communications/ Navigation Systems	USAFOMC	148	34	.12	.95
328X0	Avionic Communications Systems	USAFOMC	65	11	.41	.98
328X1	Avionic Navigation Systems	USAFOMC	83	7	.27	.97
672X2	Disbursement Accounting	AFHRL	149	20	.26	.98
304X0	Ground Radio Communications Equipment	AFHRL	335	48	.17	.98

Note. R_{11} and R_{kk} values are for the total sample (Number Raters), which includes the number of divergents ($r \leq .30$) shown.

Table 2. Training Emphasis Data Analyzed With Sequential Removal of Divergent Raters

AFSC	After First Set of Deletions: Number Raters	Number Divergent	R_{11}	R_{kk}	After Second Set of Deletions: Number Raters	Number Divergent	R_{11}	R_{kk}	Total % Deleted
404X0	35	1	.13	.84	34	2	.14	.85	28
811X0	97	2	.20	.96	95	0	.21	.96	21
328XX	114	3	.14	.95	111	2	.15	.95	25
328X0	54	0	.55	.99	-	-	-	-	17
328X1	76	2	.29	.97	74	0	.32	.97	11
672X2	129	2	.36	.99	127	0	.37	.99	15
304X4	287	4	.20	.99	283	6	.20	.99	16

Note. R_{11} and R_{kk} are for the Number of Raters, which includes the number of newly identified divergent raters ($r \leq .30$) shown.

Table 3. Variation in R_{11} with Sample Size

Sample Size	R_{11} for AFSC 672X2: $\bar{X}$	SD	MIN	MAX	R_{11} for AFSC 304X4: $\bar{X}$	SD	MIN	MAX
10	.238	.112	.017	.517	.156	.061	.025	.205
50	.257	.033	.144	.335	.167	.020	.119	.214
100	.259	.021	.211	.308	.165	.012	.132	.196
	N=129	R_{11}=.2596			N=287	R_{11}=.1686		

Note. Data elements ($\bar{X}$, SD, MIN, MAX) are for 100 randomly drawn samples for each sample size.

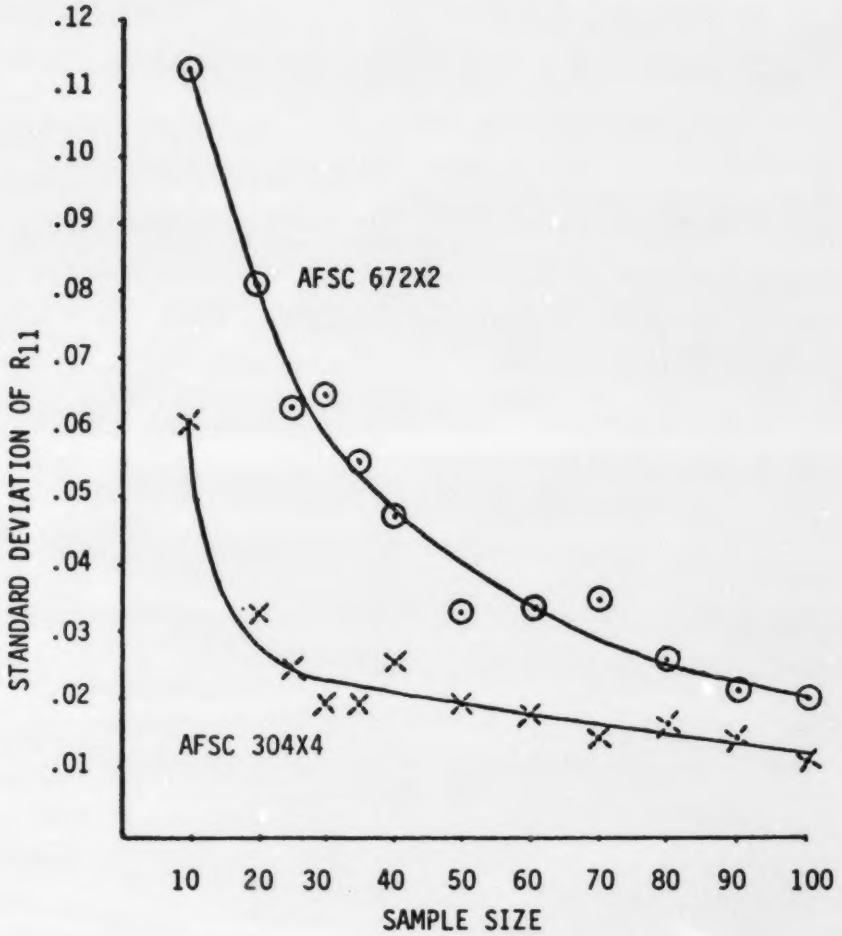

Figure 2. Stability of R_{11} versus sample size.

Detecting Multiple Rating Policies

Modified REXALL Analysis

Given that REXALL is specifically designed to evaluate rater performance with respect to a single rating policy, employing it as a tool to assist with the identification of multiple rating policies within a single data set requires that rater subgroups representing potential rating policies be somehow preselected. Modified REXALL analysis involved two different methods for predefining potential rating policy groups.

First, the possibility that a complex rating data set might be comprised of one dominant policy and a smaller minor policy was investigated by iteratively applying REXALL; i.e., by removing the raters having a relatively high correlation with the sample mean vector from the original set of raters and running REXALL on the two resulting sets of raters until stable policies and assorted divergent raters have been identified. This approach assumes that the sample mean vector is driven by the dominant policy raters and requires an arbitrary criterion correlation point to establish potential rating policy group membership. Tables 4 and 5 contain the distribution and percent occurrence of rater correlations produced by the respective sample mean vectors. A criterion correlation point of .30 to divide raters led to dominant policy

results as produced by the existing procedure for extracting the common rating policy (see Table 2). REXALL analysis of the potential minor policy groups resulted in very poor interrater agreement for all samples. Adjustment of the criterion correlation point to .40 produced very stable dominant policies for all specialties except AFSCs 404X0 and 328XX. All potential minor policy groups displayed very poor interrater agreement. Considering the arbitrary nature of the criterion correlation point, and the questionable assumption that similar rater correlations equate to similar rating patterns, the results for all samples were inconclusive with respect to confirming the presence or absence of the dominant/minor policy condition. In general, this method was a poor one for dealing with complex specialties.

Table 4. Frequency of Occurrence of Rater Correlations
(Pearson Product-Moment Correlations)

				Number of Raters Correlating with the Mean (Interval)							
AFSC	No of Raters	R_{11}	R_{kk}	1.0-.90	.89-.80	.79-.70	.69-.60	.59-.50	.49-.40	.39-.30	.29-.20
404X0	47	.09	.73				5	10	9	11	12
811X0	120	.15	.95			3	20	30	29	15	23
328XX	148	.12	.95				2	22	4	46	34
328X0	65	.41	.98	2	21	19	7	4	0	1	11
328X1	83	.27	.47		4	20	21	9	15	7	7
672X2	149	.26	.98		30	32	19	18	19	11	20
304X4	335	.17	.98			15	58	93	78	35	48

Note: The ranges are for Pearson Product-Moment Correlation Coefficients ($\underline{r}$) between individual raters and the mean rating. For example, for AFSC 404X0 there are five raters who correlate less than .7 but greater than or equal to .6 with the total sample task mean vector.

A second modified REXALL analysis method involved the analyses of potential rating policy groups comprised of raters with common background variables such as duty title, major command, and specialty code. Previously recorded high levels of interrater agreement for the two separate specialties, AFSC 328X0 and AFSC 328X1, drawn from the AFSC 328XX dual-ladder sample, constitute the only interpretable success for this method. The inconsistency of results for all other samples rendered this approach unsuitable.

Table 5. Percentage of Occurrence of Rater Correlations (Pearson Product-Moment Correlations)

AFSC	No of Raters	R_{11}	R_{kk}	Percentage of Raters		
				$r \geq .4$ Good	$.4 > r > .3$ Doubtful	$r \leq .3$ Divergent
404X0	47	.09	.73	51	23	26
811X0	120	.15	.95	68	13	19
328XX	148	.12	.95	46	31	23
328X0	65	.41	.98	81	2	17
328X1	83	.27	.47	84	8	8
672X2	149	.26	.98	80	7	13
304X4	335	.17	.98	76	10	14

Note: Percentage distribution of all REXALL rater correlations with respect to three categories: good raters ($r \geq .4$); doubtful raters ($.4 > r > .3$); and divergent raters ($r \leq .3$)

Cluster Analysis

The CODAP clustering programs were applied to the samples in an attempt to develop new procedures and guidelines for using and interpreting existing clustering software with task factor data. Appendix A provides a description of the clustering programs, the similarity measure (percent training emphasis in common), and auxiliary CODAP programs used to interpret the clusterings. For all samples, the percent-training-emphasis-overlap algorithm aggregated the raters, who were very homogeneous with respect to the number and type (by duty) of tasks rated. REXALL analysis of these main rater groups produced significantly higher values of R_{11} and higher individual rater correlations with their respective group task mean vectors than were observed with the parent sample. This indicated that those raters who have high overlap with one another on the ratings of tasks they choose to recommend for training display a high level of overall interrater agreement. Merging of these groups resulted in rater clusters with reduced levels of interrater agreement.

Group rating policies differed to varying degrees in their rank-ordering of tasks. Within each sample, the strongest differences ($r_s < .50$) occurred between groups rating virtually all or many tasks across all duties and those rating few tasks across duties or rating tasks confined to very few duties. These rating policy groups were minor in number and size and represent raters with extreme training recommendations. Less prominent policy differences ($r_s \geq .50$) occurred between groups rating closer to the sample average number of tasks rated. Raters in these groups constituted the bulk of each sample and tended to emphasize much the same technical duties which contained a large, common core of high-training-priority tasks.

The dual-specialty AFSC 328XX sample and the small AFSC 404X0 sample clusterings exhibited individual differences not observed in the other clusterings. For the AFSC 328XX sample, 89% of raters clustered into two single-specialty groups: AFSC 328X0 or AFSC 328X1. Within each single-specialty group, rating policy correlations are highly positive ($r_s > .50$). Across specialty groups, rating policy correlations are negative. The AFSC 404X0 clustering produced three small rater groups which account for only 63% of the sample. All three group rating policies demonstrate significant differences highlighted by very low between-rating-policy rank-order correlations ($r_s < .50$). Ungrouped raters (27%) were regarded as heterogeneous, isolate raters.

A valuable feature of the CODAP system is the capability to process rater background information. The CODAP DUVARS, PRTDIS, and PRTVAR rater data summaries in clustering (KPATH) sequence were found to be useful aids for interpreting observed REXALL interrater reliability statistics and rater correlations. The PRTVAR program can be utilized to summarize rater biographics in the KPATH clustering sequence to determine the extent of shared background characteristics within rater groups. For all single-specialty samples, rater characteristics, such as grade, major command, primary and duty specialty, and job title/work station (available only for AFSCs 672X2 and 304X4), could not be discerned to have any obvious connection with cluster groups. Application of discriminant analysis to establish the extent to which background variables predict cluster group membership failed to detect any meaningful associations. In the case of the dual-specialty (AFSC 328XX), raters clearly clustered into primary duty rating policy groups; i.e., either AFSC 328X0 or AFSC 328X1.

In summary, the CODAP clustering of training emphasis ratings produced cluster structures comprised of a number of rater groups with rating policy differences which were mainly a function of variation in the number and type of tasks and duties raters chose to recommend for training. However, four limitations are seen as major obstacles to accepting the training emphasis cluster structures as a generally suitable method for identifying multiple rating policies. First, the adjustment of ratings to a percentage of a rater's total rating sum results in the loss of important information about the level (magnitude) of assigned ratings. Second, the overall clustering is strongly driven by overlap over all non-zero-rated tasks, which detracts from common duty emphasis. Third, subjective decisions are required to determine the cluster group boundaries. Last, the status of the considerable number of isolate raters (5% to 20%) is an unknown. Because of these limitations, the clustering of training emphasis ratings is regarded as generating a rater sequence incorporating rater subsets which are useful only as a meaningful summary of rater characteristics and not representative of multiple rating policies.

Since a CODAP approach, if successful, would offer many operating conveniences, five additional approaches were tested for making use of the clustering programs. These techniques, which were based on assumptions not reported here, involved different treatments of the raw data prior to input to the CODAP clustering programs. The five data treatments were as follows: (a) direct input of the raw ratings to the OVRLAP program, bypassing the usual INPSTD percentage conversion described in Appendix A; (b) conversion of all non-zero ratings to values of 1, with all zeros left zero; (c) conversion of all non-zero ratings to values of 1, with all zero ratings ignored in the clustering programs; (d) conversion of all ratings by adding 1, producing a 1 to 10 rating scale, with no zeros in the analysis; and (e) a conversion designed to give higher weight to the higher raw ratings. In this last conversion all original non-zero ratings with a value of X were transformed to 2^{X-1}, and all zeros ignored in the clustering. In every case, these similarity measures generated much the same clustering group structure as the percent training emphasis clustering. The CODAP clustering approach was consequently discarded as a suitable analysis technique for identifying multiple rating policies.

Factor Analysis

A Q-type principal components factor analysis (MAX-FACTOR program) with a rater by rater correlation matrix input (TRICOR program using ratings on a 0-9 scale) was applied to each

training emphasis sample. With this approach, raters were treated as variables loading on factors (dimensions of common variance) which were interpreted as potential rating policies. The customary criterion factor loading of .33 (approximately 11% of a rater's variance accounted for) was taken as the minimum absolute value for meaningful rater contribution to a factor rating policy. Each factor rating policy was defined by examining the pattern of rater loadings in relation to considerations such as rater background characteristics, percent training emphasis per duty allocation or the rank-ordered task means for a factor rating policy group. The relative strength of rating policies was determined by comparing their respective common variances as proportions of total variance accounted for (%N).

In contrast to cluster analysis, where rating policies are characteristic of rater groups with mutually exclusive membership, factor analysis generates rating policies that are external to the rater set by determining each rater's loading on each rating policy extracted. This permits evaluation of rater performance across all policies. A further feature of this approach is the capability to control the number of rating policies for analysis. Initially, the extent to which a single general factor common rating policy prevails was investigated. By employing a VARIMAX rotation/factor building methodology, the relative utility of factor solutions consisting of iteratively increasing numbers of rating policies was evaluated in order to establish the multiple rating policy structure which best characterizes the sample and also to establish the relationship between that structure and the CRP.

General factor solution. The general factor extracted in a one-factor solution accounts for the greatest amount of shared variance within the data and is conceptualized as the CRP underlying the total rater set. Analysis of the pattern of rater loadings on this factor establishes the extent to which the CRP exists within the sample. All single-specialty samples were found to have a factor CRP characterized by all significant loadings being unidirectional and by an acceptable level of rater agreement. Except for AFSC 404X0, the common rating policy accounted for the majority of raters. In contrast, the dual-specialty AFSC 328XX general factor was comprised of bipolar significant loadings indicative of two strong specialty-specific rating policies and preclusive of a CRP as the dominant policy for the total sample. Statistics and details for this factor CRP for the single-specialty samples are presented in Table 6.

Table 6. Analysis Results for the General Factor (CRP) for Each Specialty

AFSC	Number[a] Raters	Number[a] Divergents	% Total Variance	R_{11}	R_{kk}
404X0	22	25 (53%)[b]	17.6	.22	.86
811X0	93	27 (23%)	23.5	.22	.96
328X0	54	11 (17%)	52.1	.54	.99
328X1	74	9 (11%)	37.8	.32	.97
672X2	125	24 (16%)	40.5	.38	.99
304X4	276	59 (18%)	25.5	.21	.99

[a]Number of Raters equates to number of loadings greater than criterion minimum of .33 (11% of variance).

[b]Parentheses contain number of divergents as percentage of total sample.

A detailed analysis of the high-low rater loading sequence on the single-specialty general factors confirmed the notion that this factor represents the dominant theme which links the majority of raters within the single-specialty samples. Iterative removal of raters from the low loading end of the rank-ordered general factor loading sequence resulted in a steady increase in R_{11} and R_{kk} despite decreasing sample size. This continual improvement of interrater reliability is a function of the systematic reduction of error variance and establishes the general factor loading sequence as an accurate distribution of rater performance with respect to the CRP.

Comparison of the REXALL high-low rater correlation sequence (as produced by the sample task mean vector) with the corresponding general factor high-low rater loading sequence for each single specialty revealed a close matching in rater rank-orders and correlation/loading values which tended to virtual equivalence with increasing total sample R_{11}. Corresponding factor CRP and REXALL analysis results are presented in Table 7. Except for AFSC 404X0, the CRP extraction criteria for both analysis procedures identified similar or identical divergent rater sets. Minor differences are due to the retention of a few REXALL doubtful raters ($.30 < r < .40$) the inclusion (or exclusion) of whom can be demonstrated to generate negligible perturbations in the rating policy task mean rank-order. For these five single-specialty samples, the REXALL grand task mean vector performed adequately as a standard for determining the relative worth of all raters with respect to the CRP. Large discrepancies between the factor and REXALL analyses statistics for AFSC 404X0 were caused by the relatively large number of divergent raters (53%) who did not identify significantly with the specialty CRP. Consequently, the sample task mean vector produced a REXALL rater correlation sequence which did not reflect the relative worth of raters with respect to the CRP. For this type of complex sample, routine REXALL analysis procedures are inappropriate.

Table 7. Comparison of General Factor (CRP) and Second Iteration Deletion Statistics for Each Specialty

AFSC	Number of Raters		R_{11}		R_{kk}		% Deleted	
	Factor	REXALL	Factor	REXALL	Factor	REXALL	Factor	REXALL
404X0	22	34	.22	.14	.86	.85	53	28
811X0	93	95	.22	.21	.96	.96	23	21
328X0	54	54	.54	.54	.99	.99	17	17
328X1	74	74	.32	.32	.97	.97	11	11
672X1	125	127	.38	.37	.99	.99	16	15
304X4	276	283	.21	.20	.99	.99	18	16

Note: R_{11} and R_{kk} are for Number of Raters surviving deletion; i.e., general factor CRP comprised of raters with loadings $\geq .33$ and REXALL results for raters with correlations $> .30$ after two deletion passes.

Although factor analysis was intended primarily to deal with the identification of multiple rating policies, the information conveyed by the one-factor solution, together with the factor/REXALL analyses comparisons, permits modification of the original REXALL CRP extraction criteria described in Section I of this report. In general terms, these findings demonstrate that for single-specialty samples, the reliable CRP is derived via REXALL analysis when a level of $R_{jj} \geq .20$ and $R_{kk} \geq .90$ is attained by the successive deletion of sets of divergent raters ($\underline{r} < .30$), providing R_{jj} increases with each deletion pass and no more than 25% to 30% of the sample is deleted. Allowing for the deletion of this maximum number of divergent raters and taking into account the R_{jj} stability/sample size findings, it was found that a minimum sample size of 55 raters was required to attain minimum acceptable interrater agreement. For smaller samples dictated by rater availability, $R_{jj} \geq .20$ and $R_{kk} \geq .80$ would be acceptable.

Rotated factor solutions. The VARIMAX rotation redistributes rater variance in an attempt to isolate the number of discrete rating policies that best characterizes the data in a meaningful training sense. Theoretically, a principal components analysis requires as many factors (rating policies) as there are variables (raters). The analysis produces them in order of decreasing proportions of total variance accounted for. However, it is obvious that the number of useful policies must be considerably less than the number of raters. The factor-building approach, whereby an iteratively increasing number of factors are extracted and rotated, starting with the two-factor solution, is based on the belief that, if significant multiple rating policies with potential training application exist, they should be represented by those initial factors which account for a high percentage of the total variance (%N) after rotation. Ideally, these factor rater groups would (a) display mutually exclusive membership, (b) account for most raters (with loadings greater than the criterion minimum of .33), and (c) espouse significantly different rating policies ($\underline{r}_s < .50$). More specifically, the analysis is truncated at that optimal utility point beyond which factors are dropped for interpretive purposes because they (a) consist of few or no significant loadings, (b) account for relatively small amounts of variance, (c) provide no further gains with respect to increasing the mutual exclusive membership of prior main factors, and (d) demonstrate no potential training application.

Application of the VARIMAX rotation/factor-building technique to all samples identified different rating policies ($\underline{r}_s < .50$) in two instances: the complex single specialty, AFSC 404X0, and the dual-specialty sample, AFSC 328XX. For all other samples, the rotated solution analyses reinforced the CRP as the dominant rating policy by identifying two or three main internal rating themes as minor variations of the CRP.

The three-factor solution for AFSC 404X0 appeared to be optimal. Factor group membership was mutually exclusive and accounted for 80% of the sample. Divergent raters who were not accounted for did not share significant variance beyond the three-factor solution. Statistics for the single- and three-factor solutions, together with details for the associated rating policies, are provided in Table 8. Pairwise correlation coefficients (Spearman's $\underline{r}_s$) among the three factors (3F1, 3F2, and 3F3) were low: 3F1/3F2 had $\underline{r}_s = .103$, 3F1/3F3 had $\underline{r}_s = .074$, and 3F2/3F3 had $\underline{r}_s = .305$. These values indicate significant high-priority task/duty differences (see Table 8). The rater policy groups were identified by the predominant duties they performed: (a) photographic processing and support equipment, (b) camera and audiovisual maintenance, and (c) camera maintenance. In summary, the AFSC 404X0 sample is comprised of three discrete and significantly different rating policies, one of which duplicates a very weak CRP. When combined, these competing multiple policies render the total sample complex and unsuitable for REXALL analysis. Details of the three-factor solution for AFSC 404X0 are given in Table 9.

Table 8. General and Rated Factor Statistics for AFSC 404X0

Solution	Factor Group	No. of Raters	% Total Variance	R_{11}	R_{kk}	No. of High-Priority Tasks	No. of High Priority Tasks by Duty E	F	G	H	I	J	K	L	M
General Factor	CRP	22	17.6	.22	.86	139	11	35	56	24	0	0	0	0	13
	3F1	16	16.8	.32	.91	148	11	34	64	29	0	0	0	0	10
Rotated Factors	3F2	13	10.9	.22	.78	130	9	8	9	6	41	16	15	20	6
	3F3	9	9.7	.03	.23	40	7	1	0	0	20	8	4	0	0

Notes: Factor group membership is determined by the number of loadings greater than or equal to the criterion minimum of .33. Group rating policies are described in terms of duty emphases associated with high training priority tasks identified by the FACPRT program. High-priority tasks are defined as those tasks with a mean rating greater than or equal to one standard deviation above the mean of task means. The frequency distributions of rating policy task means revealed that, complementary to their respective high-priority tasks, GRP 3F1 and GRP 3F2 assign zero-to-low training emphasis to approximately 80% of all tasks whereas GRP 3F3 allocates an average to above-average training emphasis to 95% of all tasks.

Table 9. Rotated Factor Solution for AFSC 404X0

Factor Group	Number Raters	R_{11}	R_{kk}	Rating Policy
3F1	16	.32	.91	Photographic Processing and Support Equipment
3F2	13	.33	.78	Camera and Audiovisual Maintenance
3F3	9	.03	.23	Camera Maintenance

Details for the optimal three-factor solution for the dual-specialty AFSC 328XX sample are presented in Table 10. The two main factor groups, 3F1 and 3F2, establish two uniquely different specialty-specific rating policies virtually identical to those extracted via the separate analysis of the two component specialties. Group 3F3 consists of raters who, by rating across all duties, formulate a minor CRP for the total sample. The mutual exclusivity of factor group membership and the low rank-order correlations between the rating policies they represent, render the total sample complex and unsuitable for REXALL analysis. The r_s values for the comparisons were 3F1/3F2, r_s = -.344; 3F1/3F3, r_s = -.088; and 3F2/3F3, r_s = .482.

The rotated solutions for the remaining five single-specialty samples share common features which disqualify the component factors as meaningful multiple rating policies. Each sample is

comprised of rating policies that are minor variations in the CRP. This is evidenced by (a) high inter-policy rank-order correlations, $r_s > .50$, (b) rank-order correlations with the CRP in the range of .70 to .99, (c) non-mutually exclusive membership, (d) high training priority tasks which are largely accounted for by the CRP high training priority tasks, and (e) rater memberships which are subsets of the CRP membership. These five single specialties are appropriately classified as simple or non-complex in that the REXALL CRP reliably subsumes the competing component rating policies.

Table 10. Rotated Factor Solution for AFSC 328XX

Factor Group	Number Raters	R_{11}	R_{11}	Rating Policy
3F1	54	.56	.99	AFSC 328X0 CRP (incl. one 328X1)
3F2	71	.33	.97	AFSC 328X1 CRP (incl. two 328X0)
3F3	16	.28	.86	AFSC 328XX CRP (eleven 328X1 and five 328X0)

III. APPLICATIONS

1. REXALL analysis incorporating the new CRP extraction criteria is appropriate for establishing the overall recommended training priority for a single-specialty sample. The REXALL configuration of a single-specialty sample likely to contain a reliable CRP is one with the following characteristics:

   a. Single-rater reliability, $R_{11} > .15$.

   b. Approximately 65% (or more) of raters with correlations, $r > .40$.

   c. Some rater correlations, $r > .70$.

2. REXALL rater correlation guidelines for retaining or rejecting raters as being reliable or divergent with respect to the CRP are as follows:

   a. If $r > .40$, reliable rater; retain.

   b. If $.30 < r < .40$, doubtful rater; analyze rating pattern before retaining or rejecting.

   c. If $r < .30$ and/or t-value < 3.0, divergent rater; reject.

3. Rating pattern analysis to support the retention or rejection of doubtful raters consists of evaluating the extent to which the following individual rater characteristics diverge from the majority rating pattern:

   a. Total number of non-zero responses.

   b. Mean rating and standard deviation on the 1 to 9 scale.

c. Distribution of non-zero ratings on the 1 to 9 scale.

d. Distribution of non-zero ratings across duty areas.

e. Distribution of percentage training emphasis across duty areas.

These rater characteristics are available from the CODAP PRTDIS (for 3a, 3b and 3c) and DUVARS (for 3d and 3e) programs. Rater sequencing can be in normal numeric input order or KPATH order. The latter sequence, which requires additional computing via the CODAP clustering programs (OVERLAP, GROUP, and KPATH), separates the rater sample into subgroups of raters with highly similar rating patterns and isolates raters with diverging rating patterns.

4. Applications of these criteria and guidelines would ensure extraction of a reliable CRP (if it exists) with a single-rater reliability $R_{11} \geq .20$. The interrater reliability for the final set of CRP raters (R_{kk}) will depend on the number of good raters surviving deletion. To maximize attainment of $R_{kk} \geq .90$, a minimum safe sample size of N = 55 is desirable. For smaller samples, an $R_{kk} \geq .80$ is acceptable.

5. Principal components factor analysis is appropriate for the analysis of complex single specialties which fail to attain acceptable interrater agreement with REXALL analysis using the new CRP extraction criteria and for multi-ladder survey data with a high potential for specialty-aligned multiple rating policies. The number and type (unidirectional or bipolar) of significant loadings on the one general factor solution will define the extent to which a CRP exists for a sample. Application of the VARIMAX rotation/factor-building analysis technique will determine the extent to which competing multiple rating policies exist within the sample.

6. In seeking a multiple factor solution, factor extraction and rotation should be stopped when the factors identified are found to satisfy the following guidelines:

a. High proportion of total variance accounted for.

b. Most raters are accounted for (loadings $\geq .33$) while remaining divergent raters (loadings $< .33$) are few and not included within the main factor structure.

c. Results remain relatively stable upon further extraction.

d. The policies found appear reasonable, with potential for generating coherent training strategies.

7. The veracity of a rotated solution reflecting intended rater training recommendations is directly proportional to the level of single-rater reliability (R_{11}) within each policy and to the extent that interpretable differentiation exists between factor policy/groups in terms of the following:

a. Mutually exclusive group membership.

b. Rank-order correlations ($r_s < .50$).

c. High training priority tasks.

d. Common background variables.

IV. CONCLUSIONS

1. Factor analyses of the six single-specialty training emphasis samples in this report, although uncovering more than one rating policy in each case, have demonstrated them to be less "complex" than anticipated. For five of these specialties, there was no practical difference (r_s .50) between the rating policies.

2. REXALL analysis employing the new CRP extraction criteria is adequate for CRP including all raters (ideal) and for CRP with divergency less than 25% (e.g., AFSCs 328X0, 328X1, 811X0, 672X2 and 304X4).

3. REXALL analysis is inadequate for the following sample types: (a) two or more competing rating policies (e.g., AFSC 404X0), (b) no main policies, and (c) multi-ladder surveys (e.g., AFSC 328XX).

4. Modified REXALL analysis and CODAP cluster analysis (normal or experimental types) are not adequate for identifying multiple rating policies.

5. The CODAP auxiliary summary programs (DUVARS, PRTDIS, PRTVAR, and FACPRT) have high utility for interpretation of REXALL and factor analyses.

6. Principal components factor analysis has a high utility for identifying the CRP and multiple rating policies.

REFERENCES

Christal, R. E., & Weissmuller, J. J. (1976). New CODAP programs for analyzing task factor information (AFHRL-TR-76-3, AD-A026 121). Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory.

Goody, K. (1976). Comprehensive occupational data analysis programs (CODAP): Use of REXALL to identify divergent raters (AFHRL-TR-76-82, AD-A034 327). Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory.

Lindquist, E. F. (1953). Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin.

Ruck, H. W., Thompson, N. A., Brown, R. H., & Stacy, W. J. (In preparation). Development of a task training emphasis scale and training priority equations. Manpower and Personnel Division, Air Force Human Resources Laboratory.

APPENDIX A: CODAP CLUSTERING DESCRIPTION

The main clustering programs are INPSTD, OVRLAP, GROUP, and DIAGRM. Initially INPSTD adjusts each rater's task ratings (0 to 9 scale) to a percentage of the sum of that raters training emphasis ratings, %TE. This adjustment standardizes all raters to a common mean of 100/NTASK. (NTASK is the total number of tasks in the inventory.) The OVRLAP program establishes a rater-by-rater similarity matrix using percent training emphasis in common (sum of linear overlap on corresponding tasks) as the measure of similarity. This matrix is collapsed by the GROUP program to form groups of raters with similar rating patterns. Each pair of raters or rater groups which merge during the grouping is given a contiguous block of (KPATH) sequence numbers. The hierarchical relationship between raters/groups can be graphically displayed via the DIAGRM program. A valuable CODAP feature is the set of auxiliary programs that can be utilized to report rater and group data summaries. Raters' training emphases, in terms of number of tasks rated (non-zero) per duty category and percentage of training emphasis per duty, are summarized in the DUVARS program printout. Rating patterns are summarized in the PRTDIS program printout which details each rater's performance on the 1 to 9 scale in terms of total number of tasks rated and mean, standard deviation and distribution of ratings. These summaries are especially relevant to group structure considerations when raters are listed in KPATH sequence. Analysis of the PRTVAR program output allows determination of the extent to which biographical and computed variables are shared by rater groups. For any selected cluster group, the JOBGRP program computes the percent training emphasis per duty summary as a general description of the group rating policy. Task-level differences between group rating policies can be highlighted by the comparison of task means across groups using the FACPRT program. Rank-order correlations between group task mean vectors, using the FACCOR program, test for rating policy differences.

☆U.S. GOVERNMENT PRINTING OFFICE: 1985 569 053 20014